GÖTEBORG STUDIES

IN EDUCATIONAL SCIENCES 148

Berit Carlstedt

Cognitive abilities – aspects of structure, process and measurement

ACTA UNIVERSITATIS GOTHOBURGENSIS

Cognitive abilities – aspects of structure, process and measurement

av

Berit Carlstedt

Fil kand

AKADEMISK AVHANDLING

som med tillstånd av samhällsvetenskapliga fakulteten vid
Göteborgs universitet för vinnande av doktorsexamen
framläggs till offentlig granskning

Fredagen den 3 november 2000 kl 10.00 i lokal D

Institutionen för pedagogik och didaktik, Göteborgs universitet,
Pedagogen, Frölundagatan 118, Mölndal

Table of content

Introduction 1

Models of the structure of intelligence – from Spearman to Carroll 2

The earlier measurement systems 9
Enlistment battery 1944 9
Enlistment battery 1947 11
Enlistment battery 1948 11
Enlistment battery 1949 13
Enlistment battery 1954 14
Enlistment battery 1959 – Enlistment battery 67 15
Enlistment battery 80 17

New theoretical and methodological influences 20
G equals Gf through Undheim and Gustafsson 20
The Nested Factor model 22
The multidimensionality aspect 24
Structural equation modeling 25

Construct validation and test development guided by the recent theoretical and methodological development 27
Summary of study A. Construct validity of the Swedish Enlistment battery. 27
Summary of study B. Swedish Enlistment Battery (SEB): Construct validity and latent variable estimation of cognitive abilities by the CAT-SEB. 30
Reanalyses of two earlier batteries. 34

Some theoretical and methodological aspects of the measurement of intelligence 38
Summary of study C. Item sequencing effects on the measurement of fluid intelligence 38

Summary of study D. Differentiation of cognitive abilities as a function of level of general intelligence. A latent variable approach. 43

Discussion 47
The model 47
Validity 48
Generalizability 49

Implications for future test development 50
The Gv factor 50
Sequential effects 50
Consequences of level differentiation 52
Future development 52

References 54

Study A

Study B

Study C

Study D

Acknowledgements

The articles presented in this thesis report work that I have done as a senior research officer at the National Defence Research Establishment and at the Department of Leadership of the Swedish National Defence College. Through their financing of the projects the Armed Forces Headquarters and the National Service Administration have had a direct interest in the research results, and in the steps for development of the Enlistment procedure that have been suggested.

Certain persons have been especially important to me in my work and studies leading to this thesis. Bertil Mårdberg who has been my workmate and knowledgeable teacher in the concrete work with the Enlistment Battery, and also inspired me to start my doctoral studies. Eva Johansson who encouraged me to continue to travel back and forth to Göteborg to attend new seminars after a longer break. My colleagues Henry Widén, Leif Carlstedt and Jens Andersson who relieved the pressure on my tasks in order to give me time to complete the summing-up of the thesis. I also thank Eva Ullstadius, creative item constructor and co-author, and Kjell Härnqvist who as a genuine expert reviewed this work at a late stage.

Jan-Eric Gustafsson my academic supervisor has continuously inspired me with his enthusiasm for my subject matter, generously shared his extensive theoretical and methodological knowledge, cooperated in the research studies, and read my manuscripts in every shape from thin drafts to productions ready to publish. Thank you.

Finally, my husband and colleague Leif who through all my work has shown pride and delight in my progress, and my children Jonas and Åsa who during the whole time period have expressed pleasure in what I have been doing.

Karlstad in September 2000

Berit Carlstedt

Introduction

The aim of this thesis is to further develop a system for measurement of cognitive abilities in young adults. The system in focus is the Enlistment Battery, which has been used for the assessment of intelligence in the Swedish military since the middle of the forties. Two main purposes are unfolded in the text. The first is to try to implement a hierarchical model of cognitive abilities for the Swedish Enlistment Battery. The second purpose is to enter deeper into certain theoretical aspects of importance for the interpretation and also for the improved measurement of general ability. The classical question of differentiation of abilities over the full range of intellectual capacity is addressed and the prospect to measure broad ability factors like general visualization and crystallized intelligence beside general ability is examined.

The thesis is based on the following four articles.

A. Carlstedt, B., & Mårdberg, B. (1993). Construct validity of the Swedish Enlistment Battery. *Scandinavian Journal of Psychology, 34*, 353-362.

B. Mårdberg, B., & Carlstedt, B. (1998). Swedish Enlistment Battery (SEB). Construct validity and latent variable estimation and profile prediction of cognitive abilities by the CAT-SEB. *International Journal of Selection and Assessment, 6*(2), 107-114.

C. Carlstedt, B., Gustafsson, J.-E., & Ullstadius, E. (2000). Item sequencing effects on the measurement of fluid intelligence. *Intelligence, 28*(2), 145-160.

D. Carlstedt, B. (submitted). Differentiation of cognitive abilities as a function of level of general intelligence. A latent variable approach.

Models of the structure of intelligence – from Spearman to Carroll

Factorially derived models of intelligence have attempted to specify the structure of mental abilities in the sense of specifying what factors exist and how they are related to each other. In the field of measuring individual differences of cognitive abilities the factor models have been the most applied models. The models have differed in primarily one important aspect, the acknowledgement of a general factor or not.

Spearman (1863-1945) was the first (1904) to empirically identify a general intelligence factor, *g*, that influenced measures of cognitive performance to a lesser or greater extent. He had observed positive correlations between tests and the highest correlations among abstract and complex tasks. He ordered the tests according to their reciprocal correlations and assumed that this hierarchy of correlations indicated the 'saturation' of each test of the general factor that was common for all the variables. Every variable was to be accounted for by two factors, g and s, a specific factor. The model was called the two-factor theory, a misleading designation, as Spearman regarded the specific factor as unique for every variable. This putative simplification of the model has probably contributed to the skepticism that it was shown during several decades. Spearman's characteristics of the g factor were described as three qualitative laws assumed to stipulate how new cognition is possible. These essential characteristics of the g factor were assumed to be apprehension of experience, eduction of relations, and the eduction of correlates (Spearman, 1927). Spearman was primarily interested in the identification of a general intelligence factor that seemed to be involved in varying degrees in most intellectual activities.

Thurstone (1887-1955) put the question of the structure of intelligence in an opposite way to Spearman. Instead of examining whether a table of correlation coefficients supported the existence of a general factor, he investigated how many ability factors he would have to postulate to account for the correlations

between the tests (Carroll, 1993). Thurstone published his first paper on multiple factor analysis 1931 and applied this method to identify primary ability factors that should be possible to interpret and should explain the covariances. In multiple factor analysis he rotated the factors into 'simple structure' with the goal to make the factors orthogonal. Each primary factor was interpreted from the content of the tests that had the strongest connection to the factor. As a result of his analyses of a battery of 56 tests administered to 240 students of age 18 (Thurstone, 1938), and of a study of 60 tests administered to 1,154 14-year old school children (Thurstone & Thurstone, 1941), he identified seven primary orthogonal factors. Those factors, V verbal, W word fluency, S space, N number, M memorizing, I inductive, and P perceptual-speed have had an immense influence on ability testing. This has been the most important way in which psychologists assess and describe the abilities of people. Test content and design of several test batteries, like for example Kit of factor-referenced cognitive tests (Ekstorm, French, Harman & Dermen, 1976) and several Swedish batteries have been guided by his model. Thurstone's interest was to clearly separate the primary factors but he later seemed willing to grant that his primary factors could be oblique and to admit the possible existence of Spearman's general factor (Carroll, 1993, p 56). "In effect" Carroll concludes, "a general factor was measured by the total scores for these batteries".

Sir Cyril **Burt** (1883-1975), who was contemporary to Spearman, was critical of the two-factor theory claimed by Spearman and conducted extensive empirical tests of a hierarchical model that beside a general factor and specific factors also contained group factors. Burt (1949) looking back on his work concludes, "At almost every stage the results seemed increasingly to confirm the broad hierarchical conception of mental organization – a series of abilities of greater or lesser range, each more or less independent of the rest, yet all included within a single unified system" (p. 105). He also provides an explanation on strict statistical grounds why he and Spearman

came to different conclusions of the number of factors from the same covariance matrices. Burt (1949, p. 106) ”accepted a residual correlation as significant when it was more than three times the probable error, Spearman insisted that it should be five times the probable error”.

Vernon (1905-1987), a follower of Burt in Britain, formulated the first clearly hierarchical model (1950) built on the analysis of among others thirteen tests given to 1,000 Army recruits. He used factor analytic techniques that made it possible to first extract the g factor, and then group factors of successively smaller breadth from the residual correlations. His model, thus, has a general factor at the top, on the next level below there are two major group factors influenced by *g*, verbal-educational (v:ed), and spatial-mechanical (k:m). The v:ed factor was depicted to dominate verbal and numerical ability, logical reasoning, attention and fluency factors. K:m was described to dominate technical and mechanical ability, educational grades in drawing and handicraft as well as spatial ability, psychomotor coordination, and even athletic skills. Carroll (1993, p. 60) concludes: ”There is good evidence, for example, for clustering of variables around higher-order verbal-educational and spatial-mechanical factors, and for domination of all these factors by some sort of general factor”. Carroll speaks as representative of the factor analysts that arrive at a hierarchical model starting from oblique primary factors and then calculates the second order factors from the correlations between them. Vernon (1973, p. 294) commented on his model that was formulated as a top-to-bottom model: ”I do not think it is correct to say that I regard, or have ever regarded, hierarchy as a psychological model. It is to me simply a convenient way for classifying test performances whereby one maximizes the variance of the most general factor first, then the major groupings, and so on to the minor factors. Thus my verbal- educational and spatial-mechanical factors do not represent mental abilities; they are the residual common variance left when one has taken out, or is holding constant, the *g* factor.”

Cattell (1905-1998), also a student of Spearman's but active in the USA, (1943) proposed the possible existence of two kinds of intelligences. A 'fluid' reflecting basic abilities in reasoning and higher mental processes, and a 'crystallized' intelligence reflecting the extent to which the individual has been able, partly on the basis of the level of fluid intelligence, to learn and profit from exposure to culture through education and other experiences. The term Crystallized was used to indicate an end product of experiences of mainly verbal, educational and acculturation activities at a certain point of an individual's life. The Gf-Gc theory was tested by Horn in his doctoral dissertation, and refined by Horn and Cattell (1966). In that study they, besides Gf and Gc, defined three other abilities at the same 'general' level, General visualization (Gv), General speediness (Gs) and general fluency (F). In the definition of Gv, primary perceptual aspects were included like width of visual field and depth perception, but also speeded visualization of movements, transformations of spatial patterns, maintaining orientation of objects in space, unifying disparate elements and locating a given configuration in a visual field. The Gf factor included reasoning in tasks requiring abstraction, concept formation and attainment, and the perception and eduction of relations. It will be measured best in culture fair or in novel tasks, and when it is required to retain elements in short-term memory. The Gc factor, being on the same general level, does indicate the breadth of awareness and refinement of relations previously attained, like in tasks requiring recognition or recall of such relations. In contrast to Gf, Gc will be measured most purely in tasks in which the subjects must use the previously attained concepts and relations of the "collective intelligence of a culture" (Horn & Cattell, 1966, p. 255). The factor analyses were done on the results of 45 tests (summed to form estimates of 23 primary factors) of 297 "adults-in-general", and resulted in the second order factors mentioned. The analysis was not brought further in spite of the fact that Horn and Cattell (1966, p. 267) reported that positive manifold (positive correlations) existed among the six factors, and "the

main general factors isolated in this study are not completely independent, and that a more general integrating principle operates among them." However, they sum up by saying that one of the personality factors that was included in the study - positive self-image - may be that general principle. Carroll (1993) concludes on this model, "among available models it appears to offer the most well-founded and reasonable approach to an acceptable theory of the structure of cognitive abilities" (p.62). However, in contradiction to that utterance he expresses a major reservation about the Cattell-Horn model in that it does not provide for a *g* factor to account for the correlations among the second-order factors.

In 1993 **Carroll** (1916-) published his book "Human cognitive abilities", which reports a very inclusive survey of factor-analytic studies published during the time period between 1925 and 1987. He chose 461 of approximately 1500 studies that reported factor-analyzed tests of cognitive ability. Criteria for his selection of the studies were broad samplings of variables, adequacy of design, and that his sample would be international in scope. The main part (76 %) had the US as the country of origin. England, Norway and Sweden contributed 36 studies altogether. Carroll practiced exploratory factor analysis, starting from the correlation matrices of the tests. He did his analyses blindly, not knowing which the variables were. Thus, he had to strictly rely on statistical criteria for the judgement of where to stop the analysis on each step of his bottom-to-top analysis. Principal component analysis and rotations of primary factors to simple structure was done. The best solution of factors on the lowest level (first stratum) was identified. The factors were expected to be oblique so in the next step those factors were analyzed to form factors on the higher level (second stratum). If there still were covariances between the factors at the second stratum, those were factor analyzed to form a general factor at the third stratum. A general factor was identified in about eight percent of the analyses. Higher order factors were defined from the lower order factors that in turn were defined from the content of the tests from which

they were defined. Carroll's motive for using exploratory factor analysis was that the variables would be able to "speak for themselves" and the covariances would suggest the most probable factor-analytic model. He claims that confirmatory factor analysis is appropriate only when specific hypotheses about factor structure are to be tested, and that different models can give the same fit to data.

Carroll summarizes his study in formulating a three-stratum theory with a general intelligence factor at stratum III, eight broad ability factors at stratum II, and under each one of these 4-12 narrow ability factors at stratum I. Graphically, he locates the stratum II factors in different distances from the general factor – the distance indicating their closeness to the general factor. The four closest are in order Fluid intelligence, Crystallized intelligence, General memory and learning, and Broad Visual perception. So in spite of the relatively few analyses that resulted in a general factor, Carroll stipulates the existence of it at the highest level of his model.

Carroll (1993) defined the factors from the tests that were part of the different batteries that he analyzed. He describes his criteria for classifying a factor as a general factor. It should have substantial loadings for lower-order factors or variables in several different domains, be identified on the third level in his bottom-up analyses, have high loadings for the Induction factor and low for psychomotor factors. The Gf factor (which he places on the second stratum) involves difficult tasks of induction, reasoning, problem solving and visual perception. In the reasoning area a high number of tests were classified into three types of tasks: deductive reasoning tasks, inductive tasks and quantitative reasoning tasks. These three types of reasoning tasks were typically correlated and Carroll assumes that this is largely due to the effects of higher-order factors (Gf, g). The first order factors that most often had their highest loadings on the Gc factor were verbal ability, language development, reading comprehension, but also general information, ideational fluency, spelling and numerical facility. Carroll (1993, p. 599) agrees to the crystallized intelli-

gence concept of Cattell in that ”it is a type of broad mental ability that develops through ‘investment’ of general intelligence into learning through education and experience”.

The factors Carroll classifies as Gv were primarily visualization and spatial relations, but also mechanical knowledge. He sees Gv as ”a general ability to deal with visual forms, particularly those that would be characterized as figural or geometric, and whose perception or mental manipulation is complex and difficult” (p. 609).

After this outline of the most influential factor models of the British and American traditions I will proceed the next section by describing the earlier Enlistment batteries.

The earlier measurement systems

In this section the versions of 1944 to 1980 of the Enlistment batteries are described. Besides depicting their concrete test content I have attempted to find out what theoretical influences there have been, what analytical tools have been applied, how test theory has evolved, and what types of measurement- and assessment problems have been of primary interest. Generally, little of the development work concerning the Enlistment batteries has been published. However, the first and the last decade of the history of the Enlistment battery, the 1940's and the 1990's, differ from the rest of the period in that more documentation has been available. The recent development is treated in the articles A and B.

Enlistment battery 1944

The measurement of individual differences in intellectual capacity in young men who were examined for compulsory military service was first started in 1944. The measurement took place in order to avoid that the enlistment board should have to make their judgements about the possibilities to train the conscript in the short time of ocular inspection that they had at their disposal. Some means of assistance to get an opinion of every man's mental capacity or his "general ability" was needed (Husén, 1944). Knowing this would make it possible to distribute the individuals to different branches of the military so that no contingent would be oversized according to intelligence, and to recognize those who were capable of accomplishing the more qualified jobs. To identify those who as a consequence of mental insufficiencies would be of only limited usefulness or impossible to train was another purpose with such a measurement system. Enlistment battery 1944 was developed with the American "The general classification test" (Bingham, 1942) serving as a model. Husén (1944) describes the American test that contained synonyms, arithmetic problems, and cube counting. The admi-

nistration of the tests was done as a ”spiral omnibus scale, in which the different tests followed after one another in a continuously rising level of difficulty and every third test is of the same type” (Husén, 1944, p. 117). Whether the testing was interrupted when the ability level of the subject was satisfactorily estimated does not show from the description. But, as the test was group administered and evaluated by means of manual templates, this does not seem to be an early adaptive test system. The Enlistment battery 1944 came to be composed of eight tests that together were assumed to measure ”the general intelligence – or the G factor with Spearman’s terminology” (Ekman, 1944, p.118). The tests were Opposites, Synonyms, Analogies, Number series, Ebbinghaus’s sentence completion, Yerke’s cubes, Porteus’s labyrinths and Minnesota form board. The evaluation was done as a general intelligence measure where all test results contributed to a composite score.

The subjects were then classified into five qualification groups neutrally named A – E (the American model used qualification groups of grade I – V). Ekman (1944) tested 115 conscripts with this battery and factor analyzed the results with Thurstone’s method of “repeated analysis”. He found a first general factor with the highest loadings in Opposites and Synonyms, and then in Number series. The intention was to measure the general factor. ”The test should in no way be allowed to have the character of a proficiency test, it had to examine the aptitude and not the education” (Ekman, 1944, p. 118). The two tests Cubes and Labyrinths were the core in a spatial factor, but those tests were regarded as of no use, as their loadings on the general factor were low. Judgements of the conscripts’ intellectual ability made by their commanders were included in the same analysis. Its loading on the general factor was significant but lower than that of most of the tests. Ekman points out that a synonym test carried out in five minutes gives a considerably more precise information about the subject’s general intelligence than the opinion from his superior, even if he has seen him in training during two months.

Summary: The general factor was clearly emphasized. To measure general intelligence was the main aim of the Enlistment battery. Factor analysis was used. The test developers got their inspiration for the design of the battery from American group test batteries used in the military from World War I and later.

Enlistment battery 1947

During the years 1944 and 1947 four of the tests were exchanged and replaced with new ones with higher validity i.e., ”such tests that best measured the intellectual aspects that were of interest” (Husén, 1948a, p.28). Looking at the tests that were chosen indicates that the general factor was essential. The new tests were Matrices, Concept discrimination (a pictorial version), Instructions and a Form board test. Opposites, Synonyms, Number series, and Sentence completion were retained and together the eight tests comprised the Enlistment battery 1947. Husén (1948a) used results from this battery to estimate the size of the ”ability-reserve”, to evaluate the grading system of the Swedish elementary school, and for his twin studies (Husén, 1959). This indicates that he regarded the battery as a valid measure of general intelligence. Weaknesses in these early versions were that they were highly speeded which made the tests not only differentiate according to the ability to solve the problems, but also to a great extent according to the speed at which the problem solving took place. The time for instructions of each test was included in the time available for the test, a matter that further accentuated the demand of reading speed and speed of understanding the instructions (Husén, 1950).

Summary: The general factor was strengthened through the addition of Matrices, Instructions and Concept discrimination.

Enlistment battery 1948

A considerable amount of developmental work preceded the formation of Enlistment battery 1948 (Husén, 1948b; Husén &

Henriksson, 1951). Twenty-seven tests were administered to 305 conscripts, systematically sampled according to the test results obtained at enlistment, to represent the whole ability scale. The tests were chosen guided by Thurstone's (1938) studies presented in "Primary mental abilities" and according to the group factors he had extracted (v= verbal ability, w= verbal fluency, n= numeric ability, p= perception, s= spatial ability, i= induction and m= memory). Each group factor was generally represented by three tests. Normalized results (mean=10 and sd=2) were used as measurement variables. Husén (1948b) presents the results of the factor analysis (Thurstone's multiple factor analysis method). Six factors were obtained, interpreted as a general, a spatial, a numeric, a speed factor, a reproductive, and a memory factor.

The researchers applied several criteria for the choice of tests to be included in the Enlistment battery 1948. Predictive validity was estimated from the commanding officers' assessments of general aptitude for military service and of general ability (intelligence) as criteria. Test reliability and construct validity (according to factor structure) were other conditions for the selection of the four new tests to be part of the battery. Those were Synonyms, Concept discrimination (now revised to present words instead of pictures), Number series and Matrices. All tests fulfilled the predictive validity criteria. "The sum of a test's correlation with all other tests was regarded as a rough indicator of the test's loading on the general factor" (Husén, 1948b, p. 28). This criterion was applied when the tests were chosen. Number series were chosen in order to also measure inductive-numeric ability. The Matrices test (obviously inspired by the Raven Progressive Matrices) was chosen to measure relations and correlates thinking. Another reason for the choice of this test was that it was regarded as knowledge-free since no letters, words or numerals were used for the problems. Compared to earlier versions of the battery that had been too verbally accentuated (Husén, 1950) "it was desired to, to a greater extent get a manifestation for the conscripts' thinking without the help of verbal material" (p. 4). The composite score of correctly solved

items over all tests (a sum of 160 items) seems to have given a good appraisal of g. Some important principles for the shaping of the new Enlistment battery were established. Items of the same type should be administered in separate tests, the test instructions should be separated from the problems of the test, instructions should be standardized, and multiple-choice responses were to be used. The time restrictions should not be too tight, ”in order not to discriminate to any greater extent” (Husén, 1948b, p. 33-35).

Summary: Thurstone’s primary factors were the basis for the selection of tests in the development work. The general factor was, however, the main object for the assessment of intellectual abilities. Internal criteria (factor analysis) and external criteria were used in the selection of tests for the battery.

Husén (1948b) clearly expressed the aim of the Enlistment battery to measure global ability and argued that a wide spectrum of test items should be used to accomplish this. He refers to Spearman who on the grounds of his theory about the structure of intelligence had endeavored to design tests that were as g-saturated as possible, and to how the student of Spearman, Raven, in 1938 had published his Progressive Matrices in the same tradition. With Thurstone’s multiple factor analysis Husén (1948b) finds that ”test construction has been released from the subjective judgements of the meaning of the tests, and from the one-sided dependence of the external, and in most cases very unreliable, validity criteria” (p. 7). The internal criteria of factor structure became possible to apply in the selection of tests for the battery.

Enlistment Battery 1949

One experience from the application of Enlistment battery 1948 was that the young men with the lowest ability had serious problems to understand the instructions of Number series (Husén, 1950). For the Enlistment battery of 1949, Number series were replaced with Instructions, in which the solution of each item is to be found within the verbal instruction of it. The principal idea of

this test was then used until the beginning of the 90's. Enlistment battery 1949 had in all 154 items and was evaluated as a composite score over all tests and as an IQ value. The conscripts were classified into five classes as before.

Summary: General ability was as strongly stressed as earlier. The composite score over all tests contributes to realize this measurement. Validity was presented as differences in test results in groups living in differently urbanized places (those living in the country were regarded as slower and less clever than the young men living in cities), and as correlations with school grades and with military training results.

Enlistment battery 1954

The Enlistment battery had earlier "concerned a classification of the conscripts according to their general capacity to carry through the military training. The increasing specialization of the military training makes it desirable to, in addition to examining general ability, also assess some more specific ability factors, among which, technical comprehension and numerical ability should be included" (Central Värnpliktsbyrån, 1954, p. 10). Instructions and Concept discrimination were kept from the former version because of their predictive validity to training criteria of different Army branches. The new tests should measure Technical comprehension and Numerical ability. Seven technical tests, of which several were rather spatial than purely technical and three numerical tests were tried out and evaluated according to a number of criteria. These criteria were the predictive validity of the test, its reliability, its factor structure, test result differences between various groups of conscripts and its correlation with school grades (Personalprövningsdetaljen, 1953). The tests that best fulfilled the formulated criteria were Levers, Technical comprehension and Multiplication. The Enlistment battery should now make possible a more differentiated classification, and the objective to measure only general ability was now abandoned. The evaluation of the test result as a composite score and the IQ

score was rejected at the same time. Each one of the five tests was evaluated as a normalized standard-nine scale. These five standard-nine values were summed up and the result was transformed into a new standard-nine scale and labeled ”Provgrupp”. This concept was used during the coming 40 years to express the result of the Enlistment battery and of the general ability of the conscripts.

Summary: The standardization of each test facilitated and encouraged a more differentiated interpretation of the Enlistment battery results. A graphical presentation of the test results as a profile over the tests was introduced. The general factor was still important during this period, but the purpose of measuring more specialized abilities was present as well. Correlations of the tests with criteria from military training, and factor analysis results were used in the selection of tests.

Enlistment battery 1959 – Enlistment battery 67

Instructions, Concept discrimination and Technical comprehension were retained for the Enlistment battery 1959. Multiplication was excluded because it was of no use for the classification of conscripts to different jobs (Militärpsykologiska Institutet, 1958). Levers, having only two response alternatives, tempted the subjects to guess, and was exchanged for Paper form board (evaluated in the Husén & Henricson (1951) study). With only minor changes like for example the introduction of optically readable answer sheets, these four tests were to become the basis of the Enlistment battery until 1980. Militärpsykologiska Institutet (1964) presents the meaning of the Enlistment battery result: ”The Provgrupp is an expression of the conscript’s general ability to profit by the military training” (p. 8). Furthermore it is stated that Instructions and Concept discrimination may be judged as the general verbal ability of the conscript – ”above all important for the capacity to assimilate the theoretical parts of the training” (p. 8). The Paper form board and Technical comprehension test results indicate together the ”suitability for technical and mecha-

nical training". The fact that the conscripts often had not completed any vocational education before enlistment, possible to use for the classification to different jobs, is described as a problem. The proportion of technical jobs in the military had increased. By this time a larger proportion of the population went to longer general school education than earlier and the percentage of vocationally skilled young men at the age of 18 to 20 years had decreased. This in turn put new demands on the Enlistment battery. As much as one third of the items of the battery were Technical comprehension items. Agrell (1958, p. 21) writes: "there is an acute shortage of mechanically and technically educated young men. We have to trace and take advantage of all conscripts with aptitude and knowledge in these matters". The aim of the Enlistment battery was now closer linked to the prediction of military training and the aim to measure general ability that was prominent during the first ten to twelve years had become less important.

During five years in the 60's job-analyses of all conscript positions in the Swedish military were performed, and physical and psychological requirements (standards) were expressed for all conscript positions (Centrala värnpliktsbyrån, 1968). In addition to the requirements of general intellectual ability and suitability as a conscript officer that were applied even earlier for the conscript non-commissioned officers, new requirements were added for all conscript positions. Those requirements were suitability as a soldier, certain physical and health demands, requirements of certain occupational experience, school education and certain skills for most positions. It seems that in competition with other aspects that now were of great interest to measure and satisfy in the assignment of all the conscript positions, the demands of cognitive ability and the assessment of cognitive abilities had a less conspicuous position than earlier.

Summary: This period was characterized by differentiated classification to military positions, increased interest of technical knowledge and aptitude. The theoretical influences seem negligible. Vernon for instance (1950), who was contemporary and

furthermore studied Army and Navy conscript recruits, had no visible influence on the Swedish enlistment battery and its evaluation. Other psychological assessments like suitability as a conscript officer or as a soldier were regarded as more important (see Björklund, 1961). Throughout the time period of nearly twenty years to follow a decline in theoretical interest for the measurement of cognitive abilities could be observed.

Enlistment battery 80

Enlistment battery 80 was the result of the next more extensive revision of the battery. Initially it was meant to be a parallel test to the former Enlistment battery 67. However, analyses showed that Concept discrimination and Paper form board had such shortcomings, primarily with respect to their reliability that they had to be exchanged (Ståhlberg-Carlstedt & Sköld, 1981). A Synonyms test was chosen to constitute a more precise measure of verbal ability than the earlier Concept discrimination. Metal folding after an idea by Härnqvist (1960) was chosen to improve the spatial ability assessment. The Enlistment battery should continue to be an extensive battery to match the concept of general ability. The technical component of the battery should be retained through the Technical comprehension test. The four tests of the Enlistment battery 80 (Instructions, Synonyms, Metal folding and Technical comprehension) were made equally long (40 items each). They were evaluated in the same way as earlier, as a normalized nine-point scale per test, added into a sum and then transformed into ”Provgrupp”. The job-analyses (Centrala värnpliktsbyrån, 1968) motivated the choice of a verbal test that ”relates to linguistic understanding and ability to use oral and written language”. The spatial test ”integrates the two clearest defined spatial ability components, ‘visualization’ – manipulation of objects in mind and ‘spatial relations’ – the ability to see and recognize still objects in different positions” (Ståhlberg-Carlstedt & Sköld, 1981, p. 6).

Even if the design of Enlistment battery 80 was justified by differentiated standards of ability factors in the conscript positions, the variable ”Provgrupp”, i.e., general cognitive ability, continued to be the main result of the measurement and to be used in the assignment to the conscript positions. Swedish test batteries used outside the military served as models. Such batteries were The Delta battery (Psykotekniska Institutet, 1970), DBA (Härnqvist, 1960), and Wit III (Westrin, 1967). The factor model of Thurstone (1938) was the dominant theory of the structure of intelligence. Factor tests were used without exception. The researchers stated that the factor tests should not be regarded as pure measures of a certain ability factor. The correlations between the factor tests were not explicitly interpreted as a general factor, even if the standardized sum of the four tests produced such an approximate score.

Summary: Even if the composition of the Enlistment battery was justified by the differentiated standards of the conscript positions, the ”Provgrupp” (actually the general factor) continued to be used. The requirements in cognitive ability have never been formally expressed otherwise. However, the assignment officers have probably used the test results in a more differentiated way in practical applications. The ideological climate in Sweden for the measurement of individual differences was quite disapproving from the end of the 60’s and into the beginning of the 80’s. The theoretical interest in ability testing within the organizations that were to be responsible for test development (Militärpsykologiska Institutet and the National Defense Research Establishment - SNDRE) was weak.

The development during the first years of testing of conscripts in Sweden was strongly influenced by Spearman and his concept of general ability. The general ability factor was what the test developers wanted to measure with the first batteries. Starting in the fifties and lasting more than thirty years, Thurstone’s primary factor model had a very powerful influence on the view of individual differences and of the test development

and measurement. The successor of the batteries described above, the Enlistment battery 1994 (CAT-SEB), was however, strongly guided by a modern theoretical and methodological development that had its roots in Scandinavia**.**

New theoretical and methodological influences on testing

The Swedish and international development in intelligence theory (Gustafsson, 1984, 1988; Undheim, 1981a, 1981b) and in methods for structural analysis (Jöreskog & Sörbom, 1988) took a new turn during the 1980's. In a pioneering article Gustafsson (1984) unified the models of the structure of cognitive abilities that originated from the British tradition with names such as Spearman, Burt, and Vernon and the leading American model of Cattell and Horn. The *g* factor of the British models was shown to be identical to the Fluid intelligence factor of the American model presented by Horn and Cattell (1966). This development was noticed by Bertil Mårdberg, director of research at the behavioral research department of SNDRE, who initiated the development of a new selection system, built on the recent theoretical and methodological advances. The authority that is responsible for the enlistment of conscripts, the National Service Administration, was also interested in a modern successor to Enlistment battery 80.

G equals Gf through Undheim and Gustafsson

Undheim (1981a, 1981b) like Carroll (1993) discussed the closeness of g and Gf and saw in that a possibility of a new synthesis between Spearman's g factor and Cattell's Gf-Gc model. He argues that the Gf factor actually has the status of a general factor in the sense that it is a factor general to a varied set of ability measures. However, another Scandinavian, Gustafsson (1984) made the first empirical test of a junction between the kind of models that acknowledge a g factor and those who do not, when he a few years later published his article "A unifying model for the structure of intellectual abilities". Gustafsson presented analyses of test data that indicated that Gf as it had been formulated in the Horn-Cattell model was equal to g as it had been described in the British models of Spearman and Vernon. A battery

of 13 ability tests of inductive, spatial and verbal character, and three achievement tests were administered to about 1,000 12-year-old subjects. The LISREL technique (Jöreskog & Sörbom, 1978) was used for the analyses. In a higher order model good support was obtained for primary factors of Thurstonian type. Those were in turn hypothesized to be influenced by three higher order factors - Gf, Gc and Gv. There were still covariances between those three factors and when Gustafsson introduced a general factor that was hypothesized to influence Gf, Gc, and Gv, the correlation between g and Gf was found to be 1.0. Gustafsson on the basis of those results suggested ”a three-level model (the HILI-model) with the g factor at the top, two broad factors reflecting the ability to deal with verbal and figural information, respectively, at the second-order level, and the primary factors in the Thurstone and Guilford tradition at the lowest level. It is argued that most previously suggested models are special cases of the HILI-model” (Gustafsson, 1984, p. 179).

Carroll mentions this study casually and later also applies his factor analysis methods on this data. He, however, arrives at a solution where Gv was closest to the general factor, obviously due to differences in analytic techniques – exploratory versus confirmatory FA (see Gustafsson, in press-a for an explanation).

Other studies (Undheim & Gustafsson, 1987; Gustafsson, 1989, in press-a) in which test data was tried out as higher order models have shown the same outcome. Undheim and Gustafsson (1987) performed higher-order analyses on test data from subjects of 11, 13 and 15 years and found unanimous results. In the hierarchical model of three levels, correlations of approximate unity between g and Gf were found in the three samples. In these analyses, despite that they were performed from bottom to top, the g factor was identified because its correlation with Gf was hypothesized. In the Gustafsson (1989) study, where a battery of 8 tests was administered to 207 boys from grade 6, four primary factors were identified. As the g factor was introduced at the next level, its correlation with the Induction factor was found to be 1.0. Gustafsson (in press-a) performed higher-order modeling of

the Holzinger and Swineford 24-test battery and found a standardized estimate of 1.00 for the loading of Gf on G in the model with the best statistical fit.

The Nested Factor model

Another important step was taken when Gustafsson and Balke (1993) published the article in which the Nested Factor model (NF-model) was first introduced. Gustafsson and Balke (1993) investigated the relations between cognitive test results gathered in the 6th grade and school achievement from the 9th grade regarding 866 students. An orthogonal NF model with latent variables was fitted to the results of the 16 cognitive tests. In the NF model the G factor influenced all ability test variables directly. The G factor was allowed to first capture all the variance in each test that was due to this factor. In the next step the broad factors Gc´ and Gv´ were introduced, also directly influencing tests of verbal and figural content, respectively. Those factors were ”nested” within the general factor, thus, influencing a narrower scope of tests. To notify that those factors influenced the residual variance when G-variance was captured, the prime ´ was added in the notations. Nested within those factors, narrow ability factors were influencing what subsequently remained of the variance of the test results.

An NF-model was also fitted to the 17 school-grades, resulting in orthogonal latent grade-factors as General achievement, and nested domain-specific factors like Science, Social science, Language, and Spatial-practical performance. The latent ability dimensions were then related to the latent grade-dimensions in structural equation modeling (SEM). Compared to traditional correlation analysis among variables a much more substantial pattern of relations appeared, mainly because of the definitions of the latent variables, but also because of the possibility to relate the latent variables to each other in SEM. The multidimensionality of test performance was observed and assessed by the latent ability variables and in the same way the

multidimensionality of the school-grades was observed and assessed. In consequence with their formulation of the NF model Gustafsson and Balke use the terms ”general”, ”broad”, and ”narrow” abilities for their hierarchical model. The denominations first, second and third order factors bear reference to the higher-order models. (Upper case G was introduced in the NF model to indicate that this factor was different from g in the higher order models).

The NF-models in Gustafsson and Balke (1993) allowed more straightforward interpretations of the relations between abilities and school achievement than the higher-order models did. In the HILI model and earlier hierarchical models the influence of factors of third, second, and first order on the test results had been expressed as indirect relations. Higher order models often give the impression that the higher order factors are more remote from the actual observations than the first and second order factors. What has been emphasized in the NF models is the simultaneous influence of several factors on most test results. As soon as a test is included in a primary factor (of a higher-order model) the multidimensionality aspect of the test performance is encapsulated. The important difference, however, between the factors of the NF model, is the range of observed variables that are directly influenced by the latent variables. The general factor has the broadest range – often all the variables of an ability test battery, the broad factors – e.g., Gc, Gv, Gs have a more limited scope of variables that they influence, and the narrow factors are even more limited in the scope of variables that they influence.

NF models are easier to formulate than higher order models and easier to test against empirical data. NF models are also often more parsimonious because only residual variance can be captured by narrower factors as they are introduced in the model building. So for example in a test battery including tests of induction, complex problem solving, and reasoning, no Gf factor is identified. The G factor has generally captured all systematic variance of those test results.

The effectiveness of nested factor models has been demonstrated in various studies. Rosén (1995) applied the NF latent variable model to study gender differences in ability tests and standardized achievement tests. Despite almost equal observed performance in manifest test scores between boys and girls, she found substantial differences in the latent ability dimensions. In a predictive validity study (Muthén & Gustafsson, 1996) of Armed Services Vocational Aptitude Battery (ASVAB) the NF model was applied to both latent ability factors and latent hands-on criteria factors. The structural relations between the latent dimensions showed interesting differential validity between different jobs. Regression of different composite scores on observed criteria variables showed a much less modulated pattern.

The multidimensionality aspect

Gustafsson (in press-b) postulates some implications of the recent theoretical development for the measurement of cognitive abilities in general: Tests and test items are multidimensional. This implies that several ability dimensions of general, broad and narrow kind are simultaneously influencing the observed test results. Such relations are easy to describe in the NF model by letting the different latent variables of the model influence the manifest results directly. To measure G, a broad spectrum of tests is needed, but this spectrum must contain a number of good Gf-measuring tests. This is also the way Gustafsson and Undheim (1996) advocate to achieve a G with good stability. To measure broad ability factors like Gv´ and Gc´, G must be measured at the same time in order to extract the observed variable variance captured by this factor. The narrow factors that may be extracted from the remaining variance are seldom reliable enough to have predictive power and often consist of above all test-specific variance. Gustafsson (in press-b) summarizes about the measurement implications of the hierarchical approach to intelligence: ”1. To measure constructs with high referent generality it is

necessary to use heterogeneous measurement devices. 2. A homogeneous test always measures several dimensions. 3. To measure constructs with low referent generality it is also necessary to measure constructs with high generality."

Structural equation modeling

In the articles of the thesis, confirmatory factor analysis (CFA) and structural equation modeling are applied as CFA seems to have some conclusive advantages. Bollen (1989) points to some of these advantages in contrast to exploratory factor analysis. If the scientist has some previous knowledge about the models of cognitive abilities and of what at least some of the tests measure, the model may be constructed in advance. The analyst can set the number of latent variables and define what relations between the latent and the observed variables that are supposed to be present. The relations that are hypothesized not to exist can be fixed to zero. Measurement error of each observed variable may be estimated. Obliqueness or orthogonality of the model may be hypothesized and tested by inserting covariances between the latent variables or not. The hypothesized model may be statistically tested against data, in such a way that the covariance matrix is re-created from the relations of the model in some estimation procedure (like Maximum likelihood) and the divergence of the model implied matrix from the empirical covariance matrix is determined. The divergence is often expressed as a $\chi 2$ value that is related to the number of degrees of freedom (related to the number of parameters that are estimated) of the model. Another index of model fit, RMSEA (Root mean Square Error of Approximation) (Browne & Cudeck, 1993), takes into account the model complexity. These indices are used to judge the fit of a model, or strictly speaking make the rejection of a non-fitting model possible.

If different treatments are to be evaluated, or different groups are of interest to compare, a multiple-groups model is chosen. In such a multiple-groups model all the parameters are initially

hypothesized to be equal between groups. In a model systematic successive relaxations of parameters are possible to test statistically by means of the different $\chi 2$-values in relation to the difference in degrees of freedom (Loehlin, 1987). Differences according to the equality of the structure of the relations, the strength of the relations between latent variables and observed variables, the amount of variance captured by the latent variables, the means of the latent and manifest variables etc, between such groups, may be tested very systematically.

Any hypothesized pattern of relations between latent and observed variables may be tested. The influence of the general, the broad and narrow factors on the test results could be estimated to describe the multidimensional features of cognitive test results.

Construct validation and test development guided by the recent theoretical and methodological development

The model of the structure of cognitive abilities by Gustafsson and Undheim was chosen for the developmental work that is reported in the articles of the thesis. The construct validity aspects of articles A and B are treated within this model of nested factors and with the assumption of multivariate influences on test performance. So are the topics of the articles C and D, which look into the measurement of G and into the differentiation of abilities over the different levels of general ability.

Summary of study A.
Construct validity of the Swedish Enlistment battery

Influenced by the theoretical and methodological development described above Carlstedt and Mårdberg in article A studied the Enlistment battery 80 from these new starting-points. The hierarchical model presented by Gustafsson (1984, 1988) was tested on the Enlistment battery. The construct validity of the battery was studied in confirmatory factor analyses using the LISREL program to examine the dimensions of the battery. Questions of interest were whether the battery measured G, and whether other ability dimensions like Gc and Gv could be extracted as well. Two studies were performed: the first on data collected at the regular enlistment of conscripts, and the second on a wider test battery administered at a training unit for soldiers.

In the first study three parallel samples of the population of young men coming to the enlistment were studied. The sample sizes were 501, 1058, and 1057. A hierarchical nested factor (Gustafsson & Balke, 1993) model was tested on the four tests Instructions, Synonyms, Metal folding and Technical comprehension (all scored as correctly solved odd and even items). Three consecutive models were tested on the first sample. A model with a G factor influencing all eight manifest variables was hypothesized and tested. The fit of that model was not good, indicating

that beside the G factor there was more variance to be accounted for. In the next step a residual Gv factor was assumed to influence the Metal folding and the Technical comprehension tests. The model fit improved considerably (a change of $\chi 2$/df ratio =181). However, there was still room for improvement of the model, so test-specific factors influencing three of the tests were introduced and resulted in a good model fit ($\chi 2$=25,21 df= 22). The good fit of the third model indicated that it was a good representation of the covariances between the tests, so the model was used as a reference model for the analyses of the other two samples. The model did not fit quite as well for those somewhat larger samples, but there was a considerable stability of the relations between the latent and the manifest variables of the three models. The loadings of the tests on G were high for all tests, but highest for Instructions (.84 - .90 for the three samples). No test-specific factor was identified for the Metal folding test, but the loading of this test on the Gv factor was high, indicating that this factor may have been more of a Metal folding test-specific factor than a Gv factor. With the limited scope of tests in the battery, the interpretations of the factors were generally uncertain. The G factor may have emerged as a consequence of the covariances between the tests, irrespective of its closeness to Gf. This demanded an enlarged data collection with reference tests of known Gf substance to define the G factor, and more tests of verbal and spatial character to enable the identification of a Gc factor and to get a better-defined Gv factor.

The second study was done on the test results of 113 soldiers (age 20-22) in compulsory military training, representing all levels of positions from privates to non-commissioned officers, i.e., also representing a good variation in cognitive ability. Apart from the four enlistment tests they went through three verbal tests (Opposites, Word fluency 1 and 2), three tests with recognized large Gf content (Matrices, Number series and Bongard), and three tests of spatial ability (Card rotation 1 and 2, and Figure rotation). All tests were collected from test batteries from outside the military. As the reference tests were chosen with clear

hypotheses about what factor they would load the model could be formulated and tested according to that. All tests would have loadings on the G factor. A Gc factor would influence all the verbal tests and the Gv factor all the spatial tests. The model fit was found to be good enough, subsequent to the introduction of four test-specific factors of Rotation, Word fluency, Instructions and Technical comprehension, influencing those tests.

The G factor was well established with loadings over all the tests, but highest for the Gf indicators Bongard, Number series, and Matrices. The Instructions test loaded almost equally high on G. A Gc factor could now be identified with the three new tests added. Synonyms and Opposites had the highest loadings on Gc, while the fluency tests had lower loadings. Compared to the Enlistment battery model, the loadings of Metal folding on the Gv factor decreased, and that of Technical comprehension increased, obviously because the rotation element was introduced into Gv by the Rotation tests. The loadings on G of the Synonyms test decreased, implying that the G factor obtained from the four Enlistment tests alone was too much twisted towards Gc.

To conclude, the sum of the normalized scores (Provgrupp) of the four tests of Enlistment battery 80 could be seen as a good estimate of general ability. But an even better measurement of this capacity would be obtained by the estimate of factor scores of a latent G factor of the hierarchical model. To develop the battery to measure G, and to also measure abilities like Gc and Gv, independent of G, additional tests of verbal and spatial character would have to be added and also more tests of large Gf content. ”Some of them should be as pure Gf tests as possible, but a sufficient number would have to load highly also on the Crystallized (Gc) and General visualization (Gv) factors to permit reliable differential scores” (article A, p. 361). The authors were surprised to find the low amount of variance that was due to the Gc and Gv factors and suggested that the determinacy of those factors would be too low for differential prediction.

This construct validity study laid the ground for the development of the Enlistment battery 94, which should combine

the features of a multidimensional battery, factor score evaluation principles, independent measures of G, Gc, and Gv, and computerization of the test administration. Factor scores, being more error-free predictors than composite scores, were intended to become the new assessments of the conscripts' intellectual capacity at enlistment.

The hierarchical NF model of intelligence was applied for these construct validity studies and the confirmatory factor analysis method was used in LISREL 7 (Jöreskog & Sörbom, 1988), making it possible to test and evaluate hypotheses on the factor structure of the tests.

Summary of study B.
Swedish Enlistment Battery (SEB): Construct validity and latent variable estimation of cognitive abilities by the CAT-SEB

The most apparent characteristic of the 1994 version of the Enlistment battery was the computerized testing procedure. The application of the hierarchical NF model for the construction and evaluation of the test battery was theoretically the most important and significant difference from earlier versions.

The use of the Enlistment battery within the complete procedure of enlistment of conscripts, principally 18-year old men, is described. Physical tests and medical examination are made to assess the conscripts' general health. Each conscript is interviewed and evaluated by a psychologist regarding ability to handle strenuous situations in the military. Each potential conscript officer (those 60 percent with the best cognitive ability) receives an additional evaluation regarding suitability as an officer. The purpose of the enlistment procedure is the classification of conscripts for military training in different positions. The classification is done in relation to requirement profiles concerning cognitive, personality, medical, and physiological variables in every job. Medical or psychological reasons are the bases of exemptions from military service (concerning about two percent of the population), and nearly 30 percent were placed in

the reserve to be trained when necessary. These figures have varied during different periods, mainly caused by the varying needs of conscript personnel for the military organization.

The theoretical basis of the CAT-SEB was the Scandinavian model (Gustafsson, 1984, 1988; Gustafsson & Balke, 1993; Undheim & Gustafsson, 1987) described earlier, with its general (G) factor, and orthogonal to G the broad residual factors Gc´ and Gv´. To enable the identification of a G factor close to Gf, the battery had to contain tests of non-verbal problem-solving items, of induction, and in addition to that a wide scope of other tests. To identify the Gc´ factor tests of word knowledge had to be included, and to identify the Gv´ factor spatial tests and a test of technical knowledge should be added. This implied that a larger battery of ten tests was used, but each test was shorter than those of the previous batteries were.

One might expect that the computerization of the testing should have implied completely new tests that to a great extent made use of the opportunities of the computer and thus differed from the traditional paper-and-pencil tests. Mårdberg and Carlstedt in article B (p. 109) outline three principal steps in computerized testing whereof the least radical was formulated: ”paper-and-pencil tests can be linearly adapted to computer presentation, scoring, and result evaluation can be automated”. This is close to a description of the CAT-SEB in that tests of paper-and-pencil origin were used. Further, the items were presented to each subject on the computer-screen in a predetermined order - the same for all subjects, the responses were given as mouse-clicks on multiple-choice alternatives presented as buttons on the screen. The responses and the time elapsed from the presentation of the item until the response was confirmed were stored in a database. The scoring of the tests and the evaluation of the test results into factor scores of G, Gc´, and Gv´ was of course computerized.

Quite a lot of concern was shown to the problem of different computer experience between different subjects coming to the enlistment testing. All were given a thorough introduction to the

use of the response tool – the mouse. In an evaluation of the attitudes to the CAT in general and especially to the fixed sequence of the item administration, the least talented subjects expressed satisfaction in not having to write. The most talented, however, found it frustrating that they could not go back to earlier items and not plan their solving of the items in advance, as one item at a time was presented on the screen.

During the period of development a number of different tests were tried out and evaluated regarding item characteristics and construct validity. For the implemented CAT-SEB ten tests were used to form the base for the latent factors G, Gc´, and Gv´. The battery presented the tests in the order of Synonyms 1, Block rotation (imagining a three-dimensional object from different positions), Figure series, Opposites, Technical comprehension, Groups (a figurally presented concept discrimination, or classification, test), Dice 1 (similar to Cube comparison), Metal folding (surface development), Synonyms 2, and Dice 2 (parallel to Dice 1). The number of items per test were generally 20 (range 25 – 16 items). The reliabilities varied between .85 and .70.

Article B presents the construct validity of the CAT-SEB in an NF model (Gustafsson & Balke, 1993) of the tests. In a random sample of 1,436 subjects the hypothesized influences of G, Gc´, and Gv´ on the test results were fitted to data. The sums of correct scores of odd and even items of each test were used as manifest variables, enabling also the identification of test specific factors. The model fitted data well ($\chi 2$=384.4, df = 144, RMSEA = 0.034). The results show that the G factor is measured well with the battery. All tests have high loadings on the general factor, and as expected the highest are observed for the Figure series and the Groups tests. Those tests were supposed to be the best Gf measures and their strong loadings on G support the G=Gf assumption. The Gc´ factor influenced strongly the different verbal tests and also Technical comprehension (a knowledge based test). The Gv´ factor was weaker. The loadings on the factor from the tests were moderate, and highest for the Metal

folding test. Actually, the loadings on the G factor were clearly higher.

The factor loadings reported in table 2 of article B were used to estimate factor scores of the three cognitive ability measures G, Gc´, and Gv´. The regression method of Lawley and Maxwell (1963) was applied. The factor score reliabilities were estimated as measures of ”determinacy” (Huang, 1997), i.e., the correlations between the true factor scores and the estimated factor scores. The factor determinacies were 0.95, 0.90, and 0.70, showing satisfactory reliability of G and Gc´. The weakness of the Gv´ factor is again illustrated in its low factor reliability.

It was decided to use the independent latent variable estimates of G, Gc´, and Gv´ as the cognitive ability measurements in the enlistment process. The multidimensionality of psychological tests was acknowledged in the evaluation of the test results. G must be regarded as an excellent measure of general ability and a good replacement to the former composite of normalized composites ”Provgrupp”. The three latent variable estimates would be possible to evaluate as a profile in which the different abilities could be seen related to each other. This way of evaluating the test results was contrasted to an evaluation as three composites of the inductive, verbal, and spatial test scores. Large intercorrelations between the composites were found. Again the great influence of G on the spatial tests was observed in a high intercorrelation (.85) between the inductive and the spatial composites. Lohman (1996, p. 98) discusses the relation of ”Spatial ability and g” and writes: ”tests of spatial abilities – especially performance tests that use blocks or form boards and pieces of paper that must be folded and unfolded – such tests are among the best measures of g (or Gf).” The spatial tests of CAT-SEB are exactly of the kind Lohman delineates, so there should be no surprise that they are contributing largely to G. According to Lohman (1996) the major reason why this relation is so strong is that spatial tests place extraordinary demands on working memory, which in turn is related to g (see e.g., Kyllonen & Christal, 1990). Much of the developmental work was later

concentrated on the measurement of a stronger Gv´, however with limited success, obviously because most of the experimental tests were of the same kind as those already in the battery. Spatial tests of different kinds were tried out comprising items where speeded rotation was necessary, prediction of the hit-point of moving objects was the task, perceptual speed, etc. Those tests did not fit into the Gv´ factor that was dominated by the highly G-loaded spatial tests of CAT-SEB.

The implications of the hierarchical NF model of intelligence were drawn and implemented in a computer administered test battery. Latent variable modeling was applied in the development of the battery. The theoretical background of the concept of intelligence played a more conspicuous part in the test development than for several decades.

In this context I found it interesting to view two of the old Enlistment batteries through the spectacles of the recent theoretical and methodological development in fitting the hierarchical NF model to two data sets.

Reanalyses of two earlier batteries

Husén and Henriksson (1951) presented the correlation matrix of 27 tests administered to 305 conscripts. A hierarchical NF model was tested in a CFA approach on the data with LISREL 8.30 (Jöreskog, Sörbom, duToit & duToit, 1999) within STREAMS 2.1 (Gustafsson & Stahl, 1999). Twenty-one of the tests were included in the model: Opposites, Sentence completion, Concept discrimination (words), Incomplete words, Disarranged words, Prefixes (a word fluency test), Spelling, Addition, Multiplication, Division, Arithmetic reasoning, Paper form board 1, Levers, Transmissions, Number series, Letter series, Letter cancellation (clerical), Paper form board 2, Instructions, Information and Matrices.

A model with five orthogonal nested factors – G, Gc, Gv, Speed and Math – showed the best, although not perfect, fit to data ($\chi 2$= 359.77, df = 159, p < .00 RMSEA = .064.). The model

with the relations between the factors and tests is presented in table 1.

Table 1. Standardized loadings of the tests on the nested factors.

	Factor					
Test	G	Gc	Gv	Speed	Math	Error
Opposites	0.72	0.36				0.59
Sentence completion	0.67	0.45		-0.14		0.58
Concept discrimination	0.67	0.30				0.67
Incomplete words	0.65	0.38		0.25		0.61
Disarranged words	0.65	0.33		0.28		0.62
Prefixes	0.68	0.31		0.33		0.58
Spelling	0.63	0.40			0.28	0.60
Addition	0.63			0.25	0.43	0.60
Multiplication	0.67				0.54	0.51
Division	0.74				0.30	0.61
Arithmetical reasoning	0.84	0.06		-0.24	0.14	0.46
Paper form board 1	0.59		0.16			0.79
Levers	0.43		0.58			0.69
Transmissions	0.51		0.65			0.56
Number series	0.79					0.61
Letter series	0.80					0.61
Letter cancellation	0.42			0.53		0.74
Paper form board 2	0.50		0.27	0.33		0.75
Instructions	0.77	0.19				0.60
Information	0.69	0.38				0.62
Matrices	0.68		0.11			0.73

Apart from the relations between the latent and manifest variables described in the table, covariances of -0.15 were observed between the residuals of Arithmetical reasoning and Concept discrimination, and of -0.21 between the residuals of Matrices and Division.

All tests had loadings on G, the highest were Arithmetical reasoning, Letter series and Instructions, the lowest Letter cancellation, (.42), and the Spatial tests (around .50). The verbal tests, but also Information, Instructions and Arithmetical reasoning had loadings on the Gc factor. All these tests were also clearly influenced by G. Levers and Transmissions had the strongest loadings on the Gv factor, and the Paper form board tests weaker. The Matrices test was also influenced by the Gv factor. The Speed factor influenced most strongly the Letter cancellation test, which is a true speed-test, but quite a few of the other tests had loadings on Speed probably caused by narrow time limits. The Math factor has its strongest influence on the simpler rules of arithmetic but also, however, lower influence on Spelling and on the Arithmetical reasoning problems – it may probably be interpreted as a school achievement factor.

The tests that originally were chosen for the Enlistment battery 1948 (Concept discrimination, Number series, Matrices and Synonyms - the latter test had too skewed a distribution to be included in the analysis but was regarded as a better word knowledge test than Opposites) all had high loadings on G. These high loadings and the variety of test types that were chosen for the battery would have given the good g-test that Husén aimed at.

A criterion of military competence was also reported for the 305 conscripts. This variable was included in the analysis and its regression on the latent variables was studied. Three statistically significant regression coefficients were shown; positive coefficients with G and Speed and a negative with Gc. A competent soldier thus, should be intelligent, quick in action and not verbally able (maybe quiet?).

The second reanalysis concerned the correlation matrix (Militärpsykologiska Institutet, 1958) of five tests of Enlistment battery 1954. The tests were Instructions, Concept discrimination, Multiplication, Levers, and Technical comprehension. An NF model was hypothesized with the G factor influencing all tests and a Gv factor influencing the Levers and Technical comprehension tests. The model fitted data extremely well when also

Concept discrimination was allowed to load on Gv ($\chi 2$=2.83, df=2, RMSEA= .027).

Table 2. Loadings of the 1954 battery tests on the ability factors.

	Factor		
Test	G	Gv	Error
Instructions	.90		.44
Concept discrimination	.82	.09	.56
Multiplication	.58		.82
Levers	.41	.66	.62
Technical comprehension	.66	.48	.58

The structure and the factor loadings of this battery strongly resembles that of SEB (Enlistment battery 80) reported in article A. The analysis reveals a distinct G factor with its greatest contribution from Instructions, and a Gv factor influencing most strongly the spatial ability test and to a lesser extent the Technical comprehension test. The sum of the normalized scores of the five tests seems to have given a good measure of general intelligence. The residual Gv factor however appears to be rather weak.

An evident stability of the models is obtained when the NF model is tested on the batteries from the different decades. The G factor has great influence on all the cognitive tests and no residual Gf or Induction factor is identified. Provided the battery has a wide enough scope of tests both Gc and Gv factors are possible to isolate. The design of the tests are still today much the same as they were when e.g. Thurstone (1937) presented his huge battery of 56 tests, however now interpreted multi-dimensionally instead of as primary factors.

Some theoretical and methodological aspects of the measurement of intelligence

I am now leaving the descriptive part concerning former versions of the Swedish enlistment battery that has been the focus so far. The two articles that will be treated in this second part of the thesis concern more theoretical aspects of the general factor, such as the nature of G, do we measure the concept of G as a trait or as a prerequisite for learning. How does the intention to measure broad factors like Gc and Gv beside G correspond with the facts of a varying dominance of general intelligence on different levels of abilities? The CAT-SEB assumes that the same structure of abilities holds over the full range of G. How daring is such an assumption for the validity of the ability factors?

Summary of study C.
Item sequencing effects on the measurement of fluid intelligence

Recent and earlier definitions of intelligence, presented by experts in the field, typically emphasize two aspects, the ability to solve complex problems and the ability to acquire new knowledge (Gottfredson, 1997; Sternberg, 1982). In research on the construct of G, the complexity aspect has often been the most prominent. Carroll (1993) from his extensive factor analytic survey of practically all published cognitive tests concluded that G dominates factors that emphasize the level of difficulty that can be mastered in performing reasoning, induction, visualization, and language comprehension. Guttman (1954) in his radex model, achieved through multidimensional scaling, places tests with high complexity near the center and tests with lower complexity in the periphery. The centrally located tests are typical G or Gf- measuring tests like Raven, Analogies and Series. In research on elementary cognitive tasks like reaction time (RT) it has been observed that when those tasks are made more complex (through the introduction of choice RT, of more complicated response rules, etc) their correlation with general

intelligence variables will increase (Frearson & Eysenck, 1986). Kyllonen (1996) has presented empirical evidence that support a very close connection between working memory capacity and reasoning ability. An increased complexity of a task would typically put higher demands on working memory and consequently also on G. Other indications of a relation between complexity and Gf were given in experimental studies by Stankov (Stankov & Crawford, 1993; Roberts, Beh, & Stankov, 1988). They increased complexity of a task, and introduced competing tasks, and found enlarged correlations with general intelligence.

Raven is one of few test constructors who has stressed the aim to measure learning aspects in intelligence testing. Raven, Raven, and Court (1995, p. G42) write about the special progressivity of the Raven Progressive Matrices (RPM) test: ”Each problem is the ‘mother’ or the ‘source’ of a system of thought and the order in which the problems are presented provides training in the method of working”. The test is also said to provide a built-in training program and to record the ability to learn from experience. Humphreys (1979) and Raaheim (1988) both emphasize that the ability to transfer past learning and achieved experiences to new contexts of some difficulty are important aspects of the definition of intelligence.

The overwhelming indications of a relation between complexity and G formed the background of the study presented in article C. The aim was to develop tests of increased G=Gf involvement by increasing their complexity. We intended to induce this higher level of complexity by mixing items from different inductive tests. An item to be solved was not to be preceded by an item of the same kind. In addition to the difficulty of each item this would require switching between principles for solution and, thus, put higher demands on G. Three non-verbal problem solving tests were used - Groups, Series, and Bongard, constructed to measure classification, sequential reasoning, and inductive reasoning - containing in all 22 items. The sequencing of these items was varied between two treatment groups, the heterogeneous (Het) and the homogeneous (Hom) groups. The

Hom treatment implied a traditional sequencing of the items, i.e., the same kinds of items were held together and administered with increasing difficulty. The Het treatment group was presented the same items but in the mixed order of one Groups item, one Bongard item and one Series item. This sequence was then repeated until all items were administered. Both groups of subjects (conscripts at enlistment) also went through three reference tests, Instructions, Synonyms, and Metal folding of the Enlistment battery 1980 (SEB).

CFA was used in this experimental context to evaluate the outcome of the different treatments in terms of differences in factor structures and factor loadings on the latent ability variables. A nested factor two-groups model (Gustafsson & Balke, 1993) was hypothesized and tested for the two treatments with LISREL 8.14 (Jöreskog & Sörbom, 1996) within the STREAMS (Gustafsson & Stahl, 1997) modeling environment. A two-group model was first formulated to test the equality of the Het and the Hom group with respect to the abilities measured with SEB. All parameters were constrained to be equal between the groups and the fit indices implied a good fit showing that the two groups had the same structure and level of abilities.

To investigate if the Het test had produced higher loadings on the G factor an NF model including not only the reference tests, but also the experimental tests, was hypothesized. The G factor influenced all the observed variables directly. The Synonyms and Instructions tests had loadings on the Gc factor. Test specific factors were hypothesized to influence Instructions, Metal folding, Groups, Series and Bongard. Identification of the test specific factors was made possible as the correct odd and even items per test were used as manifest variables. Initially, all parameters were constrained over treatment groups. Thus, the influences from all the latent variables on all the manifest variables were set equal between groups, as were the means, the variances of the latent variables, and the error variances in manifest variables. A poor fit was achieved of this restricted model, whereby the G influence on all three item types (Groups, Series,

and Bongard) were relaxed between groups. The model fit improved significantly and indicated that there were differences between treatments as to the G-influence on the experiment tests. Inspecting the factor loadings, however, revealed a result opposite to what was expected. There was an effect of item sequencing but not in the expected direction - the Hom treatment showed the highest G influence.

The Het procedure may have created demands to keep three solving strategies in mind and continually shift between them throughout the test. A new factor (Hetspec) that influenced only the Het treatment results of the experiment tests was introduced and resulted in a better and good-enough fit. There were larger error variances for the Het test, indicating random influences or lesser homogeneity. An opposite result was obtained for the Hom treatment in as much as the test specific variances tended to be higher than for the Het treatment.

Complexity as an important aspect for the assessment of G was of course not rejected as we may have failed in our attempt to affect this aspect. The reason why the opposite occurred may however be that the learning aspect had worked more actively in the Hom sequencing of the items. Analyses on item level regarding the manifest test results and the acquired factor structure indicated effects of the sequencing of items, and especially learning effects. Despite equal results of the general ability of SEB the Hom group performed better on the manifest results of the experiment tests, indicating a greater learning opportunity. The modeling on item level showed several factors defined by items that had strong residual correlations even after the G, Gc and Metal folding and Hetspec factors were extracted. Such items were typically influenced by the acquisition of some solving strategy that helped the test taker disentangle other items.

Most problem solving tests are not explicitly (like RPM) designed to supply training throughout the test; instead such an effect is often regarded as a detriment and a source of measurement error. In article C (p. 13), however, is written: ”In spite of the fact that psychometric tests are not designed to

measure learning potential, they may to a smaller or lesser degree offer gradual training from their sequencing of items, and this opportunity to learn from experience may primarily be taken advantage of by the high-G test-takers." Such processes may have given the higher G loadings for the Hom treatment of the study. The high-G subjects learned more from the attempted items regarding principles for solution and could also to a greater extent apply what they had learned from the preceding problems. This was especially the case for items in the Hom sequence.

Whether the chosen interpretation is correct or not, the results contradict the view that test items are discrete and independent. Instead it seems that that a change of sequence of items changes the characteristics of a test. In computer-adaptive testing in its classical form (see e.g., Wainer, 1990) discrete items are used for the decision to administer the next discrete item until an estimate of the individuals ability is achieved with the required precision. This implies that different subjects are presented completely different sequences of items and that the sequencing effects are not possible to control. In article C is argued that a way of solving this problem would be to try out groups of items and use them at testing in the same order. This will however, call for something different from item-adaptive-testing.

Last, some aspects of the measurement of G are discussed. Even if the Hom test has the higher G loadings it is far from perfect, nor can any other cognitive test perfectly measure G (Gustafsson, in press-b). Because of the unavoidable test specific components of a test it is theoretically impossible for a single test to measure G. A broad spectrum of test types that maximizes the variety of test content and of cognitive operations is instead the best condition for measuring G (Gustafsson & Undheim, 1996). In this context both verbal and spatial tests, especially complex ones, have appeared to be good measures of G (Carroll, 1993; Lohman, 1996; see also the structure of the CAT-SEB in article B). The reasons for the high G loadings on verbal tests would be the impact of general ability on the acquisition of the meaning of new words and their storage in memory. The great influence of G

on spatial ability tests would mainly arise in the test situation and be caused by the demands of maintaining and transforming images in working memory.

Summary of study D. Differentiation of cognitive abilities as a function of level of general intelligence. A latent variable approach.

Differentiation of intellectual abilities is a well-known hypothesis that has been raised from time to time. The phenomenon of differentiation has in a rough distinction been given two separate meanings. One concerns differentiation originating from development, growth, and learning possible to observe in groups of different age levels or in groups of subjects who have had different treatments. The other meaning refers to differentiation that can be observed at different levels of ability in samples of subjects of the same chronological age. Garrett (1938, 1946) was the first to formally elaborate the differentiation hypothesis. His hypothesis emphasized the change of the organization of intelligence as age increases in children from a unified and general ability to a loosely organized group of abilities. Research on the differentiation effect has focused on these two directions. It has also been observed that the influence of general ability gradually decreases when training gives higher skills in such proficiencies as spatial ability (Allen, 1978), Morse code (Fleishman & Fruchter, 1960), and reading ability (Maxwell, 1972).

The aspect of differentiation over ability levels for subjects of the same chronological age was the most interesting in the context of testing at enlistment. Several studies have investigated the average correlations between test results in groups of subjects of different ability (Detterman & Daniel, 1989; Deary, Egan, Gibson, Austin, Brand & Kellaghan, 1996; Legree, Pifer & Grafton, 1996). The classification of subjects into ability groups was typically done on the basis of different cognitive tests and the relations of the other tests in the battery were investigated as

correlations between the tests or as covariances between factors. These studies all supported the differentiation hypothesis.

It was established in article D that the differentiation hypothesis has had little or no influence on practical testing and evaluation of tests. In test evaluation it is often on the contrary assumed that the structure of abilities is the same over the full range of intellectual capacity. According to the differentiation hypothesis, however, the likelihood of identifying broad or narrow ability factors is greater for high-ability subjects than for low-ability subjects.

Study D used a multivariate approach with latent variable modeling to investigate the differentiation hypothesis. In an orthogonal model where the latent variables are directly influencing the manifest variables (NF-model) it is possible to decompose the different variance parts that is accounted for by the latent variables. Gustafsson (1997, in press-b) has presented a method for this variance decomposition, which was used here.

The classification of subjects into the different ability groups was done on the basis of an extensive measurement of general ability, the G factor score of the CAT-SEB. Within the ability groups models were fitted to sets of tests from CAT-SEB measuring Gc and Gv. The differentiation effect was studied as the proportion of the total variance that the broad factor (e.g., Gc) explained at each G level. An increase in the proportion of Gc (or Gv) variance with increasing G level would be a result in favor of the differentiation hypothesis. The approach thus focused on the variances contributed by the broad ability factors. The study differed from earlier investigations of the differentiation hypothesis in that the observed variances due to broad factors were observed directly at different ability levels and not only indirectly from the amount of variance due to G over ability levels.

A sample of 14,720 young men tested at enlistment were divided into 8, 16 and 32 ability groups of equal size according to G factor score. A number of multiple-groups models were tested and evaluated with tests of change of $\chi 2/df$ ratios and the change of the RMSEA statistic between models. Separate models for Gc

(influencing Synonyms 1 and 2, Opposites and Technical comprehension) and Gv (influencing Block rotation, Metal folding, Technical comprehension and Dice 2) were tested. Initially a model with strict constraints of equality over groups was fitted and then the constraints over groups were removed in two steps in order to test the occurrence of differentiation. The two steps included first, the error variances of observed variables was relaxed and second, the variances of the broad factor (Gc or Gv) was relaxed over groups. These broad factor variances were the most crucial for the study of the differentiation hypothesis.

The proportion of variance of the broad factors over ability levels showed a weak but significant increase over G levels. This increase was however not observed up through the highest levels of G, probably caused by a lack of difficult items with good discriminative power. In order to investigate the robustness of the techniques used in selecting subjects to groups according to G level and in estimating the amount of variance accounted for by the broad factors at each such level, a simulated data set was generated. It held the same covariance structure, means and standard deviations as the real data, assuming a multivariate normal distribution, which thus has the property of homoscedasticity of variance. This implies that no differentiation was present. The simulated data material was analyzed parallel to the real data and the results were used as comparisons to the generally small effects found in the real data. Those comparisons confirmed the results further.

The results can be interpreted in terms of Anderson's minimal cognitive architecture model (Anderson, 1992) that emphasizes the role of general ability (the basic processing mechanism) for the latent capacity of the specific processors, resulting in larger individual differences in test scores. Cattell (1987) in his investment theory regards the varying investment of fluid intelligence in complex learning situations as the basis for the divergent development of Crystallized intelligence. The higher level of G indicates a larger potential outcome of Gc

and/or Gv, on the condition that a large enough quantity of activity in relevant areas is invested.

There may be some practical implications of these findings. As, according to the differentiation hypothesis, the theoretical possibilities for estimating broad ability factors (like Gc and Gv) are smaller at the lower levels of intelligence, the goal to measure those factors over the whole range of abilities might be unworkable. With this understanding we should adjust the testing procedures. Different ability groups could be presented with different combinations of tests put together in order to measure ability factors of such a degree of specificity that we could expect from the conditions (assumptions) of the differentiation hypothesis.

There is however some controversy on this matter, mostly regarding the age differentiation aspect. Carroll (1993) in his large survey analyzed data sets from different age groups and concluded that he had found little evidence to support the hypothesis that cognitive abilities become more differentiated with age. Härnqvist (1997) expressed doubt about the existence of a continuous differentiation effect in concluding from a study of differentiation of abilities for boys and girls that he does "not exclude the possibility that factor variability develops differently for different factors, genders and age levels" (p. 61).

Discussion

Through the more than half a century long history of the Swedish Enlistment battery the general intelligence factor (G) has been measured and used for the classification and selection of conscripts in different military jobs. Even if the object to assess the profile of more differentiated ability factors was expressed during periods, the functioning measure has all the same been some generic measure of the general factor. Starting with the computerized Swedish Enlistment Battery of 1994 (CAT-SEB) the opportunity to validly measure also factors of broad and narrow abilities such as Crystallized intelligence (Gc) and General visualization (Gv) was realized through the adoption of modern theoretical and methodological development originating in Scandinavia.

The model

In the article where Cattell (1943) first proposed his model of Fluid and Crystallized intelligence he had scrutinized the different models of the structure of intelligence presented until then and noticed a considerable variety. He claimed "that factor systems now require an act of psychological decision" (p. 172) and "calls for the setting up of a psychometric definitions committee to inquire as to which system offers the greatest convenience to the greatest number". Obviously he had in mind that his model would be of influence in such a committee, being a synthesis of the earlier models. I will not make any such suggestions regarding the Scandinavian hierarchical model of cognitive abilities that has been applied throughout the thesis, but would like to verify its theoretical and methodological value for the different research efforts that were reported in the articles. The model has proven to be useful in showing the construct validity of the enlistment batteries of different epochs. It has also guided decisions about what direction new test development should take as well as the subsequent evaluation of tried-out tests.

In the experiment of item sequencing the set-up as well as the analysis was feasible as a consequence of this model of intelligence. An unconventional test of the differentiation hypothesis was made possible as direct tests of the increased variance of the residual Crystallized and Visualization factors with increasing general ability levels. The analysis tools using linear structural relations were crucial in this work as they offer the opportunity to test the different hypotheses and to test changes of hypotheses statistically.

Analyzing test performance in the latent variable context acknowledges the multidimensionality of tests as well as of items. It is obvious that in the solution of test items different processes and different content may be involved parallel and cause this multidimensionality.

Validity

One important feature of the hierarchical nested factor model is that the general aspects and the specific aspects of intelligence share the prospect to become valid measures. Even if the specific factors (actually, broad and narrow) capture small variance parts they have shown interesting differential validity as related to external criteria like school grades and criteria from military training (see e.g., Gustafsson & Balke, 1993; Muthén & Gustafsson, 1996; Muthén, Hsu, Carlstedt, & Mårdberg, 1994). It seems however that it is vital that the criteria are treated in the same way – like multivariate measures, possible to decompose into latent criterion dimensions. The decisive move is that the general aspects are partialled out of the specific, bringing those out in full relief. The validity of general intelligence has been confirmed in the most varying contexts like prediction of military job performance (Ree, Earles & Teachout, 1994) and of academic performance (Brodnick & Ree, 1995), but seems when defined as in the Scandinavian model (G=Gf) to be even more distinct in that the residual variance is captured by other factors. These are in turn possible to use as valid predictors.

Generalizability

The generalizability of the findings is limited in as much as no women were included in the studies. Since the mid 90's, women have been able to enlist voluntarily as conscripts or as officers to be (since 1982), but until then solely men were tested with the Enlistment batteries. The results of female test-takers have consistently been deleted from the data sets of the empirical studies that were presented in the thesis. No attempt will be done here to present analyses of the structure of abilities for these women in order to make comparisons to the male structure. The reason for this is that the available female samples must be regarded as highly selected in many not clearly known ways.

Numbers from the last decade indicate that more than 95 percent of the young male population have attended the testing, and those who have not, were excluded because of somatic disorders and/or mental retardation. This implies that the selective effect has been nearly non-existent. Even huge samples of data from test batteries like ASVAB used for recruitment of military personnel in the USA, or admission tests for university or college education are exposed to selective effects. The samples that are within reach at the Swedish enlistment session hold also subjects of poorer capacity. All have also been presented the same test battery.

Implications for future test development

The Gv factor

The composition of CAT-SEB has turned out to produce a Gv factor of low determinacy and a large influence of G on the spatial tests of the battery. This would however be expected if the observations of e.g., Carroll (1993) and Lohman (1996) – complex spatial tests are good G factor tests – are taken into consideration. The kind of tests that were chosen to give a broad general visualization factor seems to have put such demands on the solving of the items that the Fluid intelligence correlates like working memory or the basic processing mechanism were of great influence. To be able also to obtain factors of narrower aspects of spatial ability, batteries large enough to allow the identification of such narrower visualization factors should be tried out. Those would in turn need to be examined with respect to their predictive validity.

Sequential effects

From study C it was found that a change of the sequencing of problem solving items changed the structure of the test. Items administered in a homogeneous sequence where the possibility to learn from earlier items to later items loaded higher on G. Conceptions of intelligence typically emphasize three aspects – the eduction of relations and of correlates, the ability to learn from experience and the ability to put the achieved knowledge into practice (Gottfredson, 1997; Sternberg, 1982). Subjects high in G would thus tend to be good at all this and those low in G would tend to be poorer at all these aspects. All these characteristics would work concurrently and show in the outcome of test performance; the high G subject is presented an item, he/she probably manages to solve the item because of his/her inductive ability, learns more from this activity, and has the higher ability to apply the experience on the coming items. This

in turn will increase the probability to solve new, more complex items. The different aspects of intelligence will work together to form a positive spiral on the performance. A negative spiral seems to apply to the less talented; even if an early item is solved, the test-taker probably does not learn from this experience and thus will not be able to apply the knowledge on a coming item. And if he/she learns something, the ability to utilize this knowledge is limited and the probability to solve later items will be low. In testing, the only observable outcome may be the increased variance of the test results of a group and the larger influence of G on test results.

The item sequencing study also demonstrated that the solving of one item seemed to depend on the solving of another. In adaptive testing of its most classical form the sample of items administered to one person will be different from the sample of items presented to another person. This will follow from the successive choices of items in order to measure ability with the best precision for each individual at every moment of the testing procedure. Sequencing effects will be out of control in such a system. One way to control these effects would be to present items in an adaptive test as larger item-clusters from which the prediction of the next item-cluster with the best precision could be done. Such a layout would likewise demand the identification of the dependence between item-clusters, and that the items are tried out and later administered in the same order. The number of relations will moreover be smaller and easier to grasp in such a test design.

Item sequencing effects seem most influential on problem solving tests. Such sequencing effects should not concern tests of word knowledge because a word, if it is known, just has to be reported. Dependence between the knowledge of words seems to origin from the acquisition of the words rather than from their presentation order in a test. Nothing is known about sequencing effects of visualization tests.

Consequences of level differentiation

Another type of adaptation of the testing procedure may become a consequence of the differentiation effect studied in article D. If the ability factors of lesser breadth than G are possible to identify principally on the higher levels of intellectual capacity, only those subjects should be presented the tests that are meant to measure the residual factors Gc and Gv. This would require a chain of tests that does not present the verbal and spatial tests until an estimate of G with the requested precision is obtained. Subjects with a low G result would finish the testing at that point. The demands on a good G estimate as outlined by Gustafsson (in press-b) should of course be fulfilled i.e, some good Gf tests are needed and a certain variety of test content as well.

Future development

So far the evaluation of CAT-SEB has taken into account the multidimensionality of clusters of items (tests or half tests). The sum score of each test has been weighted in a regression equation into factor scores for G, Gc and Gv. Recent research on the construct validity of 130 new vocabulary items (Ullstadius, Gustafsson, & Carlstedt, 2000) has shown multidimensionality also on the item level. Administered together with the CAT-SEB as reference tests the vocabulary items showed considerable variation in their loadings on G and Gc respectively. This indicates that vocabulary items could be combined into tests of high G content and low Gc content or into tests of especially high Gc content. A vocabulary test containing items with preferably high-G low-Gc loadings would thus constitute a quickly administered and valid estimate of G.

Up to this point the other types of items have not been analyzed according to their dimensionality, but it seems reasonable to assume that for example spatial items could be influenced not only by G and Gv, which is to be expected, but maybe also by Gc. Problem solving items may be just as multidimensional. A

manner of making even further use of this item multidimensionality for the measurement of the ability factors would be to estimate factor scores directly from the item results. In this way, there would be an opportunity to extract all valid factor variance of the items for the factor scores, or choose not to extract for example Gc content out of a spatial item. Using sum scores from shorter or longer tests will to some extent obscure the gathering of such valid information from the items.

References

Agrell, J. (1958). Den svenska militärpsykologiens organisation och arbetsuppgifter [The organization and tasks of the Swedish military psychology]. *Tidskrift i militär hälsovård, 83*, 16-28.

Allen, M. J. (1978). An empirical demonstration of the factor differentiation hypothesis. *Multivariate Behavioral Research, 13*, 63-75.

Anderson, M. (1992). *Intelligence and development. A cognitive theory.* Oxford: Blackwell.

Bingham, H. C. (1942). The Army personnel classification system. *Annals of the American Academy of Political and Social Science.*

Björklund, E. (1961). *Utveckling av metoder för att mäta psykologiska egenskaper med särskild hänsyn till krav i krigsbefattningar* [Development of methods for assessing psychological qualifications for the demands in military conscript positions]. MPI Rapport nr 8. Stockholm: Militärpsykologiska Institutet.

Bollen, K. (1989). *Structural equations with latent variables.* New York: Wiley.

Brodnick, R. J., & Ree, M. J. (1995). A structural model of academic performance, socioeconomic status, and Spearman's *g*. *Educational and Psychological Measurement, 55*(4), 583-594.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models.* Newbury Park, CA: Sage.

Burt, C. (1949). The structure of the mind: A review of the results of factor analysis. *British Journal of Educational Psychology, 19*, 100-111, 176-199

Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies.* Cambridge: University Press.

Cattell, R. B. (1943). The measurement of adult intelligence.

Psychological Bulletin, 40(3), 153-193.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action.* Amsterdam: North- Holland.

Centrala Värnpliktsbyrån (1954). *Instruktion för provledare vid psykologiska prov i samband med inskrivningsförrättningar jämte anvisningar för provens organiserande av inskrivningschef* [Instructions for test leaders at enlistment and directions for the test administration]. Lund: Håkan Ohlssons boktryckeri.

Centrala Värnpliktsbyrån (1968). *1968 års befattningsanalys-rapport till ÖB* [Job-analysis report of 1968 to the supreme commander]. Official letter 29/4.

Deary, I. J., Egan, V., Gibson, G. J., Austin, E. J., Brand, C. R., & Kellaghan, T. (1996). Intelligence and the differentiation hypothesis. *Intelligence, 23,* 105-132.

Detterman, D. K., & Daniel, M. H. (1989). Correlations of mental tests with each other and with cognitive variables are highest for the low IQ groups. *Intelligence, 13*, 349-359.

Ekman, G. (1944). Konstruktion och standardisering av 1944 års inskrivningsprov [Construction and standardization of Enlistment battery of the year 1944]. *Tidskrift för psykologi och pedagogik 1943-44,* 118-121 .

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for KIT of factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service.

Fleishman. E. A., & Fruchter. B. (1960). Factor structure and predictability of successive stages of learning Morse code. *Journal of Applied Psychology, 44* (2), 97-101.

Frearson, W. M., & Eysenck, H. J. (1986). Intelligence, reaction time (RT) and a new "odd-man-out" RT paradigm. *Personality and Individual Differences, 7,* 807-817.

Garrett, H. E. (1938). Differentiable mental traits. *Psychological Record, 2,* 259-298.

Garrett, H. E. (1946). A developmental theory of intelligence. *The American Psychologist, 1,* 372-378.

Gottfredson, L. S. (1997). Mainstream science on intelligence:

An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24* (1), 13-23.

Gustafsson, J-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8*, 179-203.

Gustafsson, J-E. (1988). Hierarchical models of individual differences in cognitive abilities. In R.J. Sternberg (Ed.), *Advances in the psychology of human intelligence*, *Vol. 4,* (pp. 35-71). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gustafsson, J-E. (1989). Broad and narrow abilities in research on learning and instruction. In R. Kanfer, P.L. Cudeck, & R. Cudeck (Eds.), *Abilities, motivation, and methodology. The Minnesota symposium on learning and individual differences* (pp. 203-237). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gustafsson, J-E. (1997). Measurement characteristics of the IEA reading literacy scales for 9- and 10-year-olds at country and individual levels. *Journal of Educational Measurement Fall 1997, 34* (3), 233-251.

Gustafsson, J-E. (in press-a). On the relation between fluid and general intelligence: A reanalysis of the Holzinger and Swineford (1939) study. *Intelligence*.

Gustafsson, J-E. (in press-b). Measurement from a hierarchical point of view. In H. Braun, D.N. Jackson, & D.E. Wiley (Eds.), *Under construction: The role of constructs in psychological and educational measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gustafsson, J-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28* (4), 407-434.

Gustafsson, J-E., & Stahl, P. A. (1997*). STREAMS User's Guide. Version 1.7 for Windows*. Mölndal, Sweden: Multivariate Ware.

Gustafsson, J-E., & Stahl, P.A. (1999). *STREAMS User´s guide. Version 2.1 for Windows*. Mölndal, Sweden: Multivariate Ware.

Gustafsson, J-E., & Undheim, J. O. (1996). Individual differences in cognitive functions. In D. C. Berliner and R. C. Calfee

(Eds.) *Handbook of educational psychology* (pp. 186-242). New York: Simon & Schuster Macmillan.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 216-257). Glencoe, IL: Free Press.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57,* 253-270.

Huang, L-C. (1997). *Predictive validity in personnel selection using latent variable models.* Unpublished doctoral dissertation. University of California, Los Angeles.

Humphreys, L. G. (1979). The construct of general intelligence. *Intelligence, 3,* 105-120.

Husén, T. (1944). Inskrivningsprovet 1944. Förutsättningar och organisation [The Induction test 1944. Requirements and organization]. *Tidskrift för psykologi och pedagogik 1943-44,* 114-118.

Husén, T. (1948a). *Begåvning och miljö* [Ability and environment]. Stockholm: Hugo Gebers Förlag.

Husén, T. (1948b). *Konstruktion och standardisering av svenska krigsmaktens inskrivningsprov. 1948 års version* [Construction and standardization of the Swedish Armed Forces' enlistment battery. The 1948 version]. Lund: Håkan Ohlssons Boktryckeri.

Husén, T. (1950). *Några data rörande svenska krigsmaktens inskrivningsprov* [Some data from the Induction test of the Swedish Armed Forces]. Lund: Håkan Ohlssons boktryckeri.

Husén, T. (1959). *Psychological twin research*. Stockholm: Almqvist & Wiksell.

Husén, T., & Henricson, S-E. (1951*). Some principles of construction of group intelligence tests for adults*. Stockholm: Almqvist& Wiksell.

Härnqvist, K. (1960). *Manual till DBA*. [Manual of differential analysis of abilities]. Stockholm: Skandinaviska Testförlaget.

Härnqvist, K. (1997). Gender and grade differences in latent ability variables. *Scandinavian Journal of Psychology, 38*,

55-62.
Jöreskog, K. G., & Sörbom, D. (1978). *LISREL IV user's guide*. Chicago, IL: International Educational Services.
Jöreskog, K. G., & Sörbom, D. (1988). *LISREL 7. A Guide to the Program and Applications. 2nd edition.* Chicago: SPSS Inc.
Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 User's reference guide. 2nd edition*. Chicago: Scientific Software.
Jöreskog, K., Sörbom, D., du Toit, S., & du Toit, M. (1999). *LISREL 8: New statistical features*. Chicago: Scientific Software.
Kyllonen, P. C. (1996). Is working memory capacity Spearman's g? In I. Dennis & P. Tapsfield (eds.), *Human abilities. Their nature and measurement* (pp. 49-75). Mahwah, NJ: Lawrence Erlbaum Associates.
Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working memory capacity?! *Intelligence, 14*, 389-433.
Lawley, D. N., & Maxwell, A. E. (1963). *Factor analysis as a statistical method*. London: Butterworths.
Legree, P. J., Pifer, M. E., & Grafton, F. C. (1996). Correlations among cognitive abilities are lower for higher ability groups. *Intelligence, 23*, 45-57.
Loehlin, J. C. (1987). *Latent variable models. An introduction to factor, path, and structural analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Lohman, D. F. (1996). Spatial ability and g. In I. Dennis & P. Tapsfield (eds.), *Human abilities. Their nature and measurement* (pp. 97-116). Mahwah, NJ: Lawrence Erlbaum Associates.
Maxwell, A. E. (1972). The WPPSI: A marked discrepancy in the correlations of the subtests for good and poor readers. *British Journal of Mathematical Statistics, 25*, 283-291.
Militärpsykologiska Institutet (1958). *Instruktion för provledare vid psykologiska prov i samband med inskrivning jämte anvisningar för provens organiserande* [Instructions for test administrators at psychological testing on enlistment of

conscripts, and directions for organizing the tests]. Lund: Håkan Ohlssons boktryckeri.

Militärpsykologiska Institutet (1964). *Instruktion för exploratörer vid inskrivningsförrättning* [Instructions for interviewers at enlistment of conscripts]. Lund: Håkan Ohlssons Boktryckeri.

Muthén, B., & Gustafsson, J-E. (1996). *ASVAB-based job performance prediction and selection. Latent variable modeling versus regression analysis*. Unpublished manuscript, School of Education, UCLA.

Muthén, B. O., Hsu, J-W., Carlstedt, B., & Mårdberg, B. (1994). *Predictive validity assessment of the Swedish military enlistment procedure using missing data and latent variable methods.* Technical report, UCLA.

Personalprövningsdetaljen (1953). *Kortfattad redogörelse för standardisering av nytt I-prov* [Short report on the standardization of a new Enlistment battery]. Manuscript, 17/11 1953 no 101.

Psykotekniska Institutet (1970). *DELTA batteriet* [The DELTA Battery]. Stockholm: Skandinaviska testförlaget.

Raaheim, K. (1988). Intelligence and task novelty. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence*, *Vol. 4,* (pp.73-97). Hillsdale, NJ: Lawrence Erlbaum Associates.

Raven, J., Raven, J. C., & Court, J. H. (1995*). Manual for Raven´s progressive matrices and vocabulary scales. General overview*. Oxford: Oxford Psychologists Press.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology, 79*(4), 518-524.

Roberts, R. D., Beh, H. C., & Stankov, L. (1988). Hick's law, competing-task performance, and intelligence. *Intelligence, 12,* 111-130.

Rosén, M. (1995). Gender differences in structure, means, and variances of hierarchically ordered ability dimensions. *Learning and Instruction, 5*, 37-62.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology, 15*, 210-293.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: Macmillan.

Stankov, L., & Crawford, J. D. (1993). Ingredients of complexity in fluid intelligence. *Learning and Individual Differences*, *5* (2), 73-111.

Sternberg, R. J. (1982). Reasoning, problem solving, and intelligence. In R. J. Sternberg (Ed.), *Handbook of Human intelligence* (pp. 225-307). Cambridge England: Cambridge University Press.

Ståhlberg-Carlstedt, B., & Sköld, P. (1981). *Inskrivningsprov 80 – testmanual* [Enlistment battery 80 – test manual]. Stockholm: FOA rapport CH55001-H7.

Thurstone, L. L. (1937). *Psychological tests for a study of mental abilities*. Chicago: University of Chicago Press.

Thurstone, L. L. (1938). *Primary mental abilities*. Psychometric monographs, 1. Chicago: University of Chicago Press.

Thurstone, L. L., & Thurstone, T. G. (1941). *Factorial studies of intelligence*. Chicago: The University of Chicago Press.

Ullstadius, E., Gustafsson, J-E., & Carlstedt, B. (2000) *Separating the influence of general and crystallized intelligence in vocabulary test items*. Manuscript submitted for publication.

Undheim, J. O. (1981a). On intelligence II: A neo-Spearman model to replace Cattell's theory of fluid and crystallized intelligence. *Scandinavian Journal of Psychology, 22*, 181-187.

Undheim, J. O. (1981b). On intelligence IV: Toward a restoration of general intelligence. *Scandinavian Journal of Psychology, 22*, 251-265.

Undheim, J. O., & Gustafsson, J-E. (1987). The hierarchical organization of cognitive abilities: Restoring general intelligence through the use of linear structural relations (LISREL). *Multivariate Behavioral Research, 22*, 149-171.

Wainer, H. (1990). *Computer adaptive testing. A primer*. Hillsdale NJ: Lawrence Erlbaum Associates, Publishers.

Vernon, P. E. (1950). *The structure of human abilities*. London: Methuen.

Vernon, P. E. (1973). Comment on Messick's paper "Multivariate models of Cognition and personality: The need for both process and structure in psychological theory and measurement" (pp. 265-303). In J. R. Royce (Ed.), *Multivariate analysis and psychological theory*. London: Academic Press Inc.

Westrin, P. A. (1967). *WIT III. Manual* [Manual of WIT III]. Stockholm: Psykologiförlaget.